assignment_07_BasakAtanu.R

atanu

2022-05-20

```
# Assignment: ASSIGNMENT 7
# Name: Basak, Atanu
# Date: 2022-05-03

## Set the working directory to the root of your DSC 520 directory
setwd("C:\\Users\\atanu\\Documents\\BellevueUniversity_MSDS\\DSC520\\Repository\\dsc520_")


## Load the `data/r4ds/heights.csv` to
heights_df <- read.csv("data\\r4ds\\heights.csv")
#head(heights_df)
# Fit a linear model
earn_lm <-  lm(earn ~ height + sex + ed + age + race, data=heights_df)

# View the summary of your model
summary(earn_lm)
```

```
##
## Call:
## lm(formula = earn ~ height + sex + ed + age + race, data = heights_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39423  -9827  -2208   6157 158723
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -41478.4    12409.4  -3.342 0.000856 ***
## height          202.5      185.6   1.091 0.275420
## sexmale       10325.6     1424.5   7.249 7.57e-13 ***
## ed             2768.4      209.9  13.190  < 2e-16 ***
## age             178.3       32.2   5.537 3.78e-08 ***
## racehispanic  -1414.3     2685.2  -0.527 0.598507
## raceother       371.0     3837.0   0.097 0.922983
## racewhite      2432.5     1723.9   1.411 0.158489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17250 on 1184 degrees of freedom
## Multiple R-squared:  0.2199, Adjusted R-squared:  0.2153
## F-statistic: 47.68 on 7 and 1184 DF,  p-value: < 2.2e-16
```

```r
age_predict_df <- heights_df[,c('age','ed','race','height','sex')]

head(age_predict_df)
```

```
##   age ed  race   height    sex
## 1  45 16 white 74.42444   male
## 2  58 16 white 65.53754 female
## 3  29 16 white 63.62920 female
## 4  91 16 other 63.10856 female
## 5  39 17 white 63.40248 female
## 6  26 15 white 64.39951 female
```

```r
predicted_df <- data.frame(
  earn = predict(earn_lm, age_predict_df),
  ed=age_predict_df$ed, race=age_predict_df$race, height=age_predict_df$height,
  age=age_predict_df$age, sex=age_predict_df$sex
  )
head(predicted_df)
```

```
##       earn ed  race   height age    sex
## 1 38666.11 16 white 74.42444  45   male
## 2 28859.09 16 white 65.53754  58 female
## 3 23301.90 16 white 63.62920  29 female
## 4 32189.84 16 other 63.10856  91 female
## 5 27807.39 17 white 63.40248  39 female
## 6 20154.60 15 white 64.39951  26 female
```

```r
## Compute deviation (i.e. residuals)
mean_earn <- mean(predicted_df$earn)
mean_earn
```

```
## [1] 23154.77
```

```r
## Corrected Sum of Squares Total
sse <- sum((fitted(earn_lm) - heights_df$earn)^2)
sse
```

```
## [1] 3.52289e+11
```

```r
ssr <- sum((fitted(earn_lm) - mean(heights_df$earn))^2)
ssr
```

```
## [1] 99302918657
```

```r
sst <- ssr + sse
sst
```

```
## [1] 451591883937
```

```r
## Corrected Sum of Squares for Model
ssm <- sum((mean_earn - predicted_df$earn)^2)
ssm
```

```
## [1] 99302918657
```

```r
## Residuals
residuals <- earn_lm$residuals
#residuals
## Sum of Squares for Error
sse <- sum((fitted(earn_lm) - heights_df$earn)^2)
sse
```

```
## [1] 3.52289e+11
```

```r
## R Squared
r_squared <- summary(earn_lm)$r.square
r_squared
```

```
## [1] 0.2198953
```

```r
## Number of observations
n <- nrow(heights_df)
n
```

```
## [1] 1192
```

```r
## Number of regression paramaters
p <- 8
## Corrected Degrees of Freedom for Model
dfm <- p-1
## Degrees of Freedom for Error
dfe <- n-p
## Corrected Degrees of Freedom Total:   DFT = n - 1
dft <- n-1

## Mean of Squares for Model:   MSM = SSM / DFM
msm <-  ssm/dfm
## Mean of Squares for Error:   MSE = SSE / DFE
mse <- sse / dfe
## Mean of Squares Total:   MST = SST / DFT
mst <- sst / dft
## F Statistic
f_score <-  msm/mse
f_score
```

```
## [1] 47.67785
```

```r
## Adjusted R Squared R2 = 1 - (1 - R2)(n - 1) / (n - p)
adjusted_r_squared <- 1 - (1 - r_squared)*(n - 1) / (n - p)
adjusted_r_squared
```

```
## [1] 0.2152832
```